The ace04-eval Scoring Formulas

There are three primary tasks in ACE 2004 – Entity Detection and Tracking (EDT), Relation Detection and Characterization (RDC), and Event Detection and Characterization (RDC). Each of these task is essentially a detection and recognition task – the target objects are *detected* in the input language stream and then the various attributes and characteristics of these objects are *recognized*.

Evaluation requires, as a preliminary step, that a correspondence (mapping) be made between the ACE system output and a reference. This mapping is chosen so that the performance measure used for system evaluation is maximized. The performance measure for all three tasks is formulated in terms of a synthetic application value, where value is accrued by correctly detecting the target objects and correctly recognizing their attributes, and where value is lost by falsely detecting target objects or incorrectly determining attributes of the target objects. The value formulas are given below:

Entity scoring

The entity evaluation score is defined to be the sum of the values of all system output entities:

$$EDT\_Value_{sys} \quad = \quad \sum_i value\_of\_sys\_entity_i$$

The value of each system output entity is defined to be the product of an inherent entity value and the sum of the values of the entity's mentions:

$$Value_{sys\_entity} \quad = \quad Entity\_Value(sys\_entity) \cdot \sum_m Mention\_Value(sys\_mention_m)$$

The *Entity_Value* of a system output entity is a function of its type. If the output entity is mapped, then the minimum value for the sys entity and its corresponding ref entity is used. For unmapped system entities, *Entity_Value* is weighted by a false alarm penalty. For mapped output entities, *Entity_Value* is discounted for errors in entity type, subtype and class:

$$Entity\_Value \quad = \quad \left\{ \begin{array}{l} \min\begin{pmatrix} ETypeValue(sys), \\ ETypeValue(ref_{sys}) \end{pmatrix} \cdot \left(W_{Eerr-type} \cdot W_{Eerr-subtype} \cdot W_{Eerr-class}\right) \text{ when mapped} \\ \\ \quad ETypeValue(sys) \cdot \left(W_{E\text{-}FA}\right) \text{ when entity not mapped} \end{array} \right\}$$

The *Mention_Value* of a system entity mention is also a function of its type. If the mention is mapped, then the minimum value for the sys mention and its corresponding ref mention is used.[1] For mapped system mentions, *Mention_Value* is discounted for errors in mention type, role and style. For unmapped system mentions[2], *Mention_Value* is weighted by a false alarm penalty and a coreference discount[3]:

$$Mention\_Value \quad = \quad \left\{ \begin{array}{l} \min\begin{pmatrix} MTypeValue(sys), \\ MTypeValue(ref_{sys}) \end{pmatrix} \cdot \left(W_{Merr-type} \cdot W_{Merr-role} \cdot W_{Merr-style}\right) \text{ when mapped} \\ \\ \quad -MTypeValue(sys) \cdot \left(W_{M-FA} \cdot W_{M-CR}\right) \text{ when mention not mapped} \end{array} \right\}$$

[1] The mapping of system output mentions to reference mentions is chosen so as to maximize the total value of the mentions.

[2] All mentions of a system output entity are unmapped for entities that are themselves unmapped.

[3] The coreference discount is intended to reduce the penalty for mentions that are valid mentions of an entity but that are incorrectly associated at the entity level. This is considered to be less harmful than mentions that are totally spurious.

For cross-document entities (i.e., for entities that are mentioned in multiple documents), the *Value* of each system entity is accumulated over all documents being evaluated.

Relation scoring

The relation evaluation score is defined to be the sum of the values of all system output relations:

$$RDC\_Value_{sys} \quad = \quad \sum_i value\_of\_sys\_relation_i$$

The value of each system output relation is defined to be the product of an inherent relation value and the sum of the values of the relation's entity arguments:

$$Value_{sys\_relation} \quad = \quad Relation\_Value(sys\_relation) \cdot \sum_a Argument\_Value(sys\_argument_a)$$

The *Relation_Value* of a system output relation is a function of its type. If the output relation is mapped, then the minimum value for the sys relation and its corresponding ref relation is used. For unmapped system relations, *Relation_Value* is weighted by a false alarm penalty. For mapped output relations, *Relation_Value* is discounted for errors in relation type and subtype:

$$Relation\_Value \quad = \quad \left\{ \begin{matrix} \min\begin{pmatrix} RTypeValue(sys), \\ RTypeValue(ref_{sys}) \end{pmatrix} \cdot \left(W_{Rerr-type} \cdot W_{Rerr-subtype}\right) \text{ when mapped} \\ \\ RTypeValue(sys) \cdot \left(W_{R\text{-}FA}\right) \text{ when relation not mapped} \end{matrix} \right\}$$

The *Argument_Value* of a system relation argument is the *Entity_Value* of that entity argument, where the entity argument of the system relation is mapped to the corresponding argument of the reference relation:[4]

$$Argument\_Value \quad = \quad Entity\_Value(sys)$$

Mapped arguments with an "unacceptably" small *Argument_Value* are assigned an *Argument_Value* of zero.[5]

For cross-document relations (i.e., for relations that are mentioned in multiple documents), the *Value* of each system relation is accumulated over all documents being evaluated. Only those argument entity mentions that appear in these documents are used to compute *Argument_Value*, however.[6]

[4] For symmetric relations, argument order is not fixed. In this case, the order used is the order which maximizes the sum of argument values is the order used.

[5] In order for a system output argument to be reasonably considered to represent its corresponding reference argument it is required to exhibit a reasonable overlap with the reference, in terms of *Entity_Value*. Specifically, the *Entity_Value* of the system output argument (mapped to its corresponding reference argument) is compared to the (self-referenced) *Entity_Value* of the corresponding reference argument. A reasonable overlap exists whenever this ratio is greater than or equal to Θ_{Amin}.

[6] The mapping of system arguments to reference arguments is done globally, however, and considers all mentions of the entity arguments. Thus the mapping, while globally optimum, may be suboptimum when considering only a single document.

Event scoring

The event evaluation score is defined to be the sum of the values of all system output events:

$$VDC\_Value_{sys} \quad = \quad \sum_i value\_of\_sys\_event_i$$

The value of each system output event is defined to be the product of an inherent event value and the sum of the values of the event's entity participants:

$$Value_{sys\_event} \quad = \quad Event\_Value(sys\_event) \cdot \sum_p Participant\_Value(sys\_participant_p)$$

The *Event_Value* of a system output event is a function of its type and its modality. If the output event is mapped, then the minimum value for the sys event and its corresponding ref event is used. For unmapped system events, *Event_Value* is weighted by a false alarm penalty. For mapped output events, *Event_Value* is discounted for errors in event type and modality:

$$Event\_Value \quad = \quad \left\{ \begin{array}{c} \min\begin{pmatrix} VTypeValue(sys) \cdot VModeValue(sys), \\ VTypeValue(ref_{sys}) \cdot VModeValue(ref_{sys}) \end{pmatrix} \cdot (W_{Verr-type} \cdot W_{Verr-mode}) \text{ when mapped} \\ \\ VTypeValue(sys) \cdot (W_{V\text{-}FA}) \text{ when not mapped} \end{array} \right\}$$

The *Participant_Value* of a system event participant is the *Entity_Value* of that entity participant, where the entity participant of the system event is mapped to the corresponding participant of the reference event.[7] For mapped participants, *Participant_Value* is discounted for errors in participant role. For unmapped system arguments, *Participant_Value* is weighted by a false alarm penalty:

$$Participant\_Value \quad = \quad \left\{ \begin{array}{l} Entity\_Value(sys) \cdot (W_{Perr\text{-}role}) \text{ when mapped} \\ \\ Entity\_Value(sys) \cdot (W_{P-FA}) \text{ when participant not mapped} \end{array} \right\}$$

Participants with zero *Participant_Value* are considered to be unmapped. Further, mapped participants with an "unacceptably" small *Participant_Value* are assigned a *Participant_Value* of zero.[8]

For cross-document events (i.e., for events that are mentioned in multiple documents), the *Value* of each system event is accumulated over all documents in which the event is mentioned. Only those event entity mentions that appear in these documents are used to compute *Participant_Value*, however.[9]

[7] The mapping of the participants of a system output event to those of a reference event is done so as to maximize the sum of the participant values.

[8] In order for a system output participant to be reasonably considered to represent its corresponding reference participant it is required to exhibit a reasonable overlap with the reference, in terms of *Entity_Value*. Specifically, the *Entity_Value* of the system output participant (mapped to its corresponding reference participant) is compared to the (self-referenced) *Entity_Value* of the corresponding reference participant. A reasonable overlap exists whenever this ratio is greater than or equal to Θ_{Pmin}.

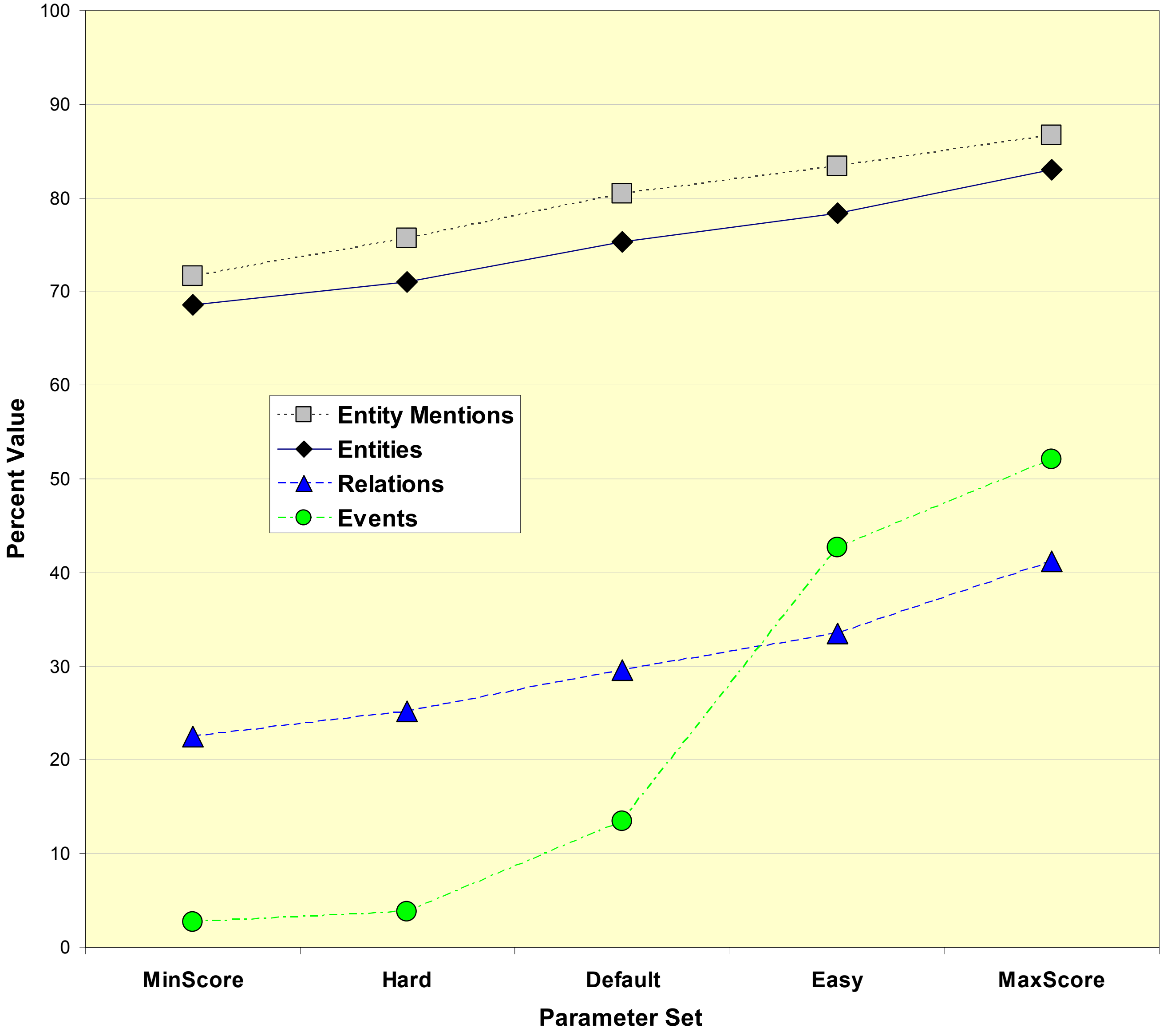
[9] The mapping of system participants to reference participants is done globally, however, and considers all mentions of the entity arguments. Thus the mapping, while globally optimum, may be suboptimum when considering only a single document.

Parameter Adjustment

The scoring parameters may be adjusted to suit the application. There are currently five sets of parameters available as command line options. In addition to the default parameters there is a set called “Easy”, a set called “Hard”, a set called “MaxSscore” and a set called “MinScore”:

	MinScore	Hard	Default	Easy	MaxScore
Entities:					
ETypeValue					
for **PER**	1.00	1.00	1.00	1.00	1.00
for **ORG/VEH/WEA**	0.50	0.50	0.50	0.50	0.50
for **GPE**	0.25	0.25	0.25	0.25	0.25
for **LOC**	0.10	0.10	0.10	0.10	0.10
for **FAC/TMP**	0.05	0.05	0.05	0.05	0.05
MTypeValue					
for **NAM**	1.00	1.00	1.00	1.00	1.00
for **NOM/BAR/MWH/PRE**	0.20	0.20	0.20	0.20	0.20
for **PRO/HLS/PTV/WHQ**	0.04	0.04	0.04	0.04	0.04
for all others	0.00	0.00	0.00	0.00	0.00
$W_{Eerr\text{-}type}$	0.00	0.00	0.50	0.75	1.00
$W_{Eerr\text{-}subtype}$	0.00	0.50	0.90	1.00	1.00
$W_{Eerr\text{-}class}$	0.00	0.50	0.75	1.00	1.00
$W_{E\text{-}FA}$	1.00	1.00	0.75	0.50	0.00
$W_{Merr\text{-}type}$	0.00	0.50	0.90	1.00	1.00
$W_{Merr\text{-}role}$	0.00	0.50	0.90	1.00	1.00
$W_{Merr\text{-}style}$	0.00	0.50	0.90	1.00	1.00
$W_{M\text{-}FA}$	1.00	1.00	0.75	0.50	0.00
$W_{M\text{-}CR}$	1.00	1.00	0.75	0.50	0.00
Relations:					
RTypeValue (for all types)	1.00	1.00	1.00	1.00	1.00
$W_{Rerr\text{-}type}$	0.00	0.00	0.50	0.75	1.00
$W_{Rerr\text{-}subtype}$	0.00	0.50	0.90	1.00	1.00
$W_{R\text{-}FA}$	1.00	1.00	0.75	0.50	0.00
Θ_{Amin}	0.75	0.50	0.25	0.00	0.00
Events:					
VTypeValue (for all types)	1.00	1.00	1.00	1.00	1.00
VModeValue (for all modalities)	1.00	1.00	1.00	1.00	1.00
$W_{Verr\text{-}type}$	0.00	0.00	0.50	0.75	1.00
$W_{Verr\text{-}mode}$	0.00	0.50	0.75	1.00	1.00
$W_{V\text{-}FA}$	1.00	1.00	0.75	0.50	0.00
$W_{Perr\text{-}role}$	0.00	0.00	0.50	0.75	1.00
$W_{P\text{-}FA}$	1.00	1.00	0.75	0.50	0.00
Θ_{Pmin}	0.75	0.50	0.25	0.00	0.00

In order to get some sense of the impact of these different sets of scoring parameters on the overall score, a small data set was processed using all four parameter sets. The ref and the sys data sets used for comparison were both produced by LDC annotators. The results are plotted below:


Percent Value
Entity Mentions
Entities
Relations
Events
MinScore
Hard
Default
Easy
MaxScore
Parameter Set